1 An overview of oral language assessment

- Introduction
- Outcome-based research on oral language assessment
  - What are language assessment interviews?
  - Past research on oral language assessment
  - The ACTFL OPI
  - Empirical studies on the OPI
  - Research on other oral examinations
  - The need for process-based research
- Discourse-based studies on oral language assessment
  - Background
  - Lazaraton's research on ESL course placement interviews
  - Participant behaviour in oral interviews
  - Comparisons of interview behaviour with conversation
  - Comparisons of test format
  - Comparisons of test scores with produced discourse
  - Rating scale construction and validation
- Conclusion

Introduction

From its historical roots in the United Kingdom in 1913, and later in 1930 in the United States, the testing of English for speakers of other languages has become what we think of as modern language testing today (see Spolsky 1990, 1995 for a detailed examination of this topic). Bachman (1991), among others, has argued that language testing as a discipline has come of age within applied linguistics, as evidenced by its achievements – its attention to theoretical issues, including theories of language ability and the effects of test method and test taker characteristics, its methodological advances in psychometrics as well as statistical analyses (see Bachman and Eignor 1997), and its impact on test development, particularly communicative testing. The language testing community now has its own refereed international journal, *Language Testing*, several international conferences, such as The Language Testing Research Colloquium (LTRC), and has published numerous books on language testing written for the teacher (e.g., Bachman and Palmer 1996; Cohen 1994; Underhill 1987) and for other language testers (e.g., Bachman 1990; McNamara 1996; and the books in this series, *Studies in Language Testing*).

However, as Bachman (1991) points out, there are other areas in language testing in which further progress is needed. For example, the interface between second language acquisition (SLA) and language testing is not as strong as it could be (see for example, Bachman 1989; Shohamy 1994a; Swain 1993; Upshur and Turner 1999; and Valdman 1988). Additionally, we have only begun to see work on the role of technology in language testing, such as computers (see Brown 1997), and speech recognition technology (as in the PhonePass™ examination, www.ordinate.com). The ethics of language testing is also a topic of current interest (see, for example, the special issue of *Language Testing*, Ethics in Language Testing, Volume 14, 3, 1997). But as far as I am concerned, the most important development in language testing over the last ten or so years is the introduction of qualitative research methodologies to design, describe, and, most importantly, to validate language tests.

In general, qualitative research has a rather short history in the field of applied linguistics, which is still trying to grapple with its legitimacy (see Edge and Richards 1998 on this point). A comprehensive overview of the methodological features of interpretive qualitative research (especially ethnography) as it is conceptualized and carried out in applied linguistics can be found in Davis (1995). Briefly, Davis discusses the important role of personal perspective in qualitative research, as well as the central focus of ‘grounded theory’, which endeavours to connect ‘a study by describing the relationships among the various parts, and it provides a theoretical model for subsequent studies’ (p. 440). Davis also discusses the issue of obtaining contextualized information from multiple data sources (triangulation) in order to achieve research credibility. Davis points out that ‘Data analysis generally involves a search for patterns of generalization across multiple sources of data … the analytic inductive method used in interpretive qualitative research allows for identification of frequently occurring events based on the data themselves. However, assertions should account for patterns found across both frequent and rare events. For assertions to hold any credibility, systematic evidence in the form of thick description must be presented in the research report’ (p. 446). According to Davis, the use of narrative, quotation from notes and interviews, and transcribed discourse from tapes are all useful in presenting results. ‘Particular description essentially serves the purpose of providing adequate evidence that the author has made a valid analysis of what the events mean from the perspectives of actors in the events' (p. 447). Davis also points out that the generalizability of data patterns can be described using frequency expressions such as 'all', 'most', 'a few', 'tended to', and 'generally', simple frequency counts, and inferential statistics.

But in a related article, Lazaraton (1995a) argues that the requirements of ethnography do not adequately account for the other ten or so qualitative research traditions in existence, traditions which have different disciplinary roots, analytic goals, and theoretical motivations. In fact, the guidelines discussed by Davis do not necessarily apply to other qualitative research approaches, particularly to qualitative discourse analysis in general, and to conversation analysis in particular.

The field of education, however, has a fairly long history of embracing qualitative research techniques, and this may account for the less skeptical reception of qualitative approaches to language testing in, for example, bilingual education. As far back as 1983, work was being done on the assessment of language minority children using ethnographic and discourse analytic techniques (see Rivera 1983). As Bennett and Slaughter (1983) note, 'The use of the analysis of discourse as a method of assessing language skills has very recently gained a high degree of respectability within the field of language proficiency assessment. The recent upsurge in interest in this area coincides with an increase in efforts to make basic research applicable to specific social problems' (p. 2). Furthermore, according to Philips (1983: 90), 'From a methodological point of view, an ethnographic perspective holds that experimental methodologies can never enable us to grasp the nature of children's communicative competence because such methods, by their very nature, alter that competence. Instead, observation, participant observation, and interviews are recommended as the research tools to be used in determining the nature of children's communicative competence.'

But it wasn't until 1984, when Cohen proposed using a specific qualitative technique, namely, introspection, to understand the testing process, that calls for a broader range of work in language testing became more frequent (e.g., Alderson, Clapham, and Wall 1995; Bachman 1990, 1991). Grotjahn (1986) warned that a reliance on statistical analyses alone will not give us a full understanding of what a test measures, that is, its construct validity; he proposed employing more introspective techniques for understanding language tests. Fulcher (1996a) observes that test designers are employing qualitative approaches more often, a positive development since 'many testing instruments do not contain a rigorous applied linguistics base, whether the underpinning be theoretical or empirical. The results of validation studies are, therefore, often trivial' (p. 228). A new respect for qualitative research as a legitimate endeavor in language testing can be seen even in unlikely places (e.g., Henning 1986 applauds the trend towards more quantitative research in applied linguistics research articles since quantitative methodology has 'certain profound advantages' over other research techniques, and yet, four years later, Dandonoli and Henning (1990: 21) remark on the 'fruitful data which can be obtained from ethnographic and qualitative research').

Specifically, more attention to and incorporation of discourse analysis in language test validation is needed (Fulcher 1987; Shohamy 1991). Fulcher remarks that 'a new approach to construct validation in which the construct can be empirically tested can be found in discourse analysis' (p. 291). Shohamy believes that tests need to elicit more discourse and to assess such language carefully, and she mentions conversation analysis specifically as one tool for examining the interaction that takes place in oral examinations. Douglas and Selinker (1992: 325) came to a similar conclusion empirically, in their study of ratings assigned to candidates taking three different oral examinations: 'This led us to a validation principle, namely that rhetorical/grammatical interlanguage analysis may be necessary to disambiguate subjective gross ratings on tests.'

McNamara (1997: 460) sees much the same need, as he states rather eloquently: 'Research in language testing cannot consist only of a further burnishing of the already shiny chrome-plated quantitative armour of the language tester with his (too often his) sophisticated statistical tools and impressive n-size'; what is needed is the 'inclusion of another kind of research on language testing of a more fundamental kind, whose aim is to make us fully aware of the nature and significance of assessment as a social act.'

The remainder of this chapter is devoted to describing the oral language assessment interview in more detail. First, a definition of an oral interview is given, followed by a summary of empirical outcome-based studies on oral assessment. The chapter concludes with a further summary of more recent discourse-based work on the interview, work which uses the actual talk produced as the basis for analysis.

Outcome-based research on oral language assessment

What are language assessment interviews?

There is some variation in terminology associated with language assessment interviews. Whereas such an encounter may be referred to as an 'oral proficiency interview', this usage can be misleading since the ACTFL OPI, the Oral Proficiency Interview, is an interview of a distinctive kind. Sometimes these assessment procedures are called 'oral interviews' or 'language interviews' as well. He and Young (1998: 10) prefer the term 'language proficiency interview' (LPI), which they define as follows:

> *'a face-to-face spoken interaction usually between two participants (although other combinations do occur), one of whom is an expert (usually a native or near-native speaker of the language in which the interview is conducted), and the other a nonnative speaker (NNS) or learner of the language as a second or foreign language. The purpose of the LPI is for the expert speaker – the interviewer – to assess the NNS's ability to speak the language in which the interview is conducted. The participants meet at a scheduled time, at a prearranged location such as a classroom or office in a school, and for a limited period. In the case of scripted interviews, an agenda specifying the topics for conversation and the activities to take place during the LPI is prepared in advance. The agenda is always known to the interviewer but not necessarily to the NNS. In addition to the agenda, the interviewer (but usually not the NNS) has access to one or more scales for rating the NNS's ability in the language of the interview.'*

The Cambridge examinations (on which much of the empirical work reported in this book is based) are referred to as Speaking Tests which employ two Examiners who rate the candidate, one an Interlocutor who conducts the assessment, and the other a passive Assessor who observes, but does not take part in the testing encounter. This terminology will be used in reference to the Cambridge examinations.

Past research on oral language assessment

The assessment of second language speaking proficiency, particularly as measured by the Foreign Service Institute–Interagency Language Roundtable (FSI/ILR) interview (Lowe 1982; Fulcher 1997: 78) considers it 'the generic ancestor of today's generation of oral tests'), the ACTFL/ETS Oral Proficiency Interview (OPI) (ACTFL 1986), and the Speaking Tests in the University of Cambridge Local Examinations Syndicate examinations (UCLES 1998c), has been a topic of considerable interest to the language testing community in the latter half of the 20th century (see Fulcher 1997 for a historical overview). There is now an extensive body of research on issues such as construct validity (e.g., Bachman and Palmer 1981, 1982; Dandonoli and Henning 1990; Henning 1992; Magnan 1988; Reed 1992), reliability and rating procedures (e.g., Bachman, Lynch and Mason 1995; Barnwell 1989; Brown 1995; Conlan, Bardsley and Martinson 1994; McNamara and Lumley 1997; Shohamy 1983; Styles 1993; Thompson 1995; Wigglesworth 1993; Wylie 1993), comparisons with other oral testing methods (e.g., Clark 1979, 1988; Clark and Hooshmand 1992; Douglas and Selinker 1992; Henning 1983; Stansfield and Kenyon 1992), aspects of the communicative competence construct (e.g., Henning and Cascallar 1992), and other aspects of oral testing (e.g., Chalhoub-Deville 1995; Clark and Lett 1988; Hill 1998; Merrylees and McDowell 1998; Raffaldini 1988; Shohamy 1988; Upshur and Turner 1999).

The ACTFL OPI

The ACTFL OPI is the most widely used face-to-face oral proficiency examination in North America, which has put it in a position to receive (perhaps more than) its fair share of criticism. For example, Lantolf and Frawley (1985, 1988) object that the ACTFL definitions of proficiency are based on intuitions rather than empirical facts about natural communication (see also Clark and Lett 1988 on this point), and on a native speaker norm which is indefensible. Bachman and Savignon (1986) and Bachman (1988) believe, first, that the OPI does not distinguish language ability from test method in its current form, thus limiting our capability to make inferences about language ability in other untested contexts, and second, that it is based on a view of unitary language ability, namely 'proficiency,' a stance which is supported by neither theory nor research. Lantolf and Frawley (1988: 10) make a similar point: 'Proficiency is derived from policy and not from science or empirical inquiry.' Kramsch (1986) takes issue with the construct of proficiency itself, pointing out that it is *not* synonymous with interactional competence. Finally, Savignon (1985) criticizes ACTFL's 'obsession with accuracy'. In response to this last point, Magnan (1988) suggests that Savignon and others have defined 'grammar' too narrowly, if not erroneously, since the skill as rated also includes appropriateness. (See also Hadley (1993) for additional responses to these criticisms of the OPI.)

But the basic objection to the OPI procedure is that is incapable of measuring what it should, namely, oral proficiency. One criticism is that the oral interview cannot provide a valid sample of other speech events because it samples a limited domain of interaction (Byrnes 1987; Clark and Lett 1988; Raffaldini 1988; Shohamy 1988). Raffaldini claims that the oral interview format, which is basically conversational, is the main reason why it fails to tap some important aspects of communication: a limited number of speech functions is sampled and so interviewees have little opportunity to display either discourse or sociolinguistic competence. Byrnes (1987: 167) admits that the ratings of the oral interview underrepresent pragmatic and sociolinguistic ability, while overemphasizing linguistic ability. But this is due to the fact that L2 studies 'rarely look at global performance features such as hesitations, false starts, repairs, and corrections', and, as a result, their meaning for aspects of communicative competence is unknown. Without this information, a description of sociointeractional, sociocultural, and sociocognitive ability cannot be included in oral proficiency rating scales.

Byrnes also makes an important point about the role of the tester in the interview. It is incumbent upon the interviewer, she maintains, to be 'keenly aware' of natural conversational behaviour, and to attempt to engage the interviewee in a 'genuine conversational exchange (the archetype occurrence of spoken language) to offset the constraints of the testing procedure' (1987: 174). This implies not only that the interview is not in itself conducive to interactional, negotiated speech, but that the achievement of a negotiated form of interaction in an interview is a collaborative accomplishment between interviewer and interviewee. To remedy this situation, Shohamy (1988) proposes a framework for testing oral language that includes a variety of interactions, each including a variety of contextual factors, that approximate 'the vernacular', which is what the oral interview fails to do. Another possibility that Clark and Lett (1988) suggest is that we check if candidates can do what the scales imply they can in the real world, perhaps by gathering self-ratings or second party ratings.

Empirical studies on the OPI

In response to these criticisms of the OPI, a number of studies have been undertaken to provide empirical evidence for the reliability and validity of this assessment procedure and the underlying ACTFL Proficiency Guidelines. For example, Dandonoli and Henning (1990; see also Henning 1992) conducted a multitrait-multimethod validation of these guidelines by considering OPI data from 60 French as a Second Language and 59 English as a Second Language students at American universities. They conclude that 'the analyses provide considerable support for the use of the Guidelines as a foundation for the development of proficiency tests and for the reliability and validity of the OPI' (p. 20).

Another validation study, focusing specifically on the role of grammar in the OPI guidelines, is Magnan's (1988) research on 40 novice-mid through advanced-plus speakers studying French. She looked at the frequency of incorrect grammatical usage of seven syntactic categories (verb conjugation, tense, determiners, adjectives, prepositions, object pronouns, and relative pronouns) to determine how they were distributed by proficiency level. She found there was a significant relationship between accuracy and level, but it was not linear and was highly dependent on the particular grammatical structure in question.

Reed (1992) looked at 70 OPIs given to ESL students at an American university in order to determine if the OPI gives 'unique' information when compared with the TOEFL. He concluded that the OPI does measure distinct skills and is thus construct valid.

Henning and Cascallar (1992) sought to determine how the four components of communicative competence (grammatical, discourse, sociolinguistic, and strategic, as per Canale and Swain (1980)), are related to each other and what their construct validity is. They tested 79 American university students on 18 performance variables, 6 pragmatic functions, 2 social registers, and 2 modalities; raters assessed 5-minute intervals of performance on a variety of communication activities. Subjects also took the TOEFL, TWE, and TSE. Among the many results were the presence of a strong interaction between performance variables and pragmatic/situational (register) functions; the importance of strategic variables in language assessment; and the continuing need to assess language structure directly, even in 'communicative' tests.

Other research has compared the face-to-face OPI with a corresponding semi-direct assessment instrument, the SOPI (Semi-Direct Oral Proficiency Interview). J. L. D. Clark has conducted several studies comparing direct and semi-direct tests. His 1979 paper discusses the methods in terms of their reliability, validity, and practicality, and concludes that semi-direct tests are 'second-order substitutes' for more direct tests (p. 48). In an empirical study, Clark (1988) compared the live and SOPI formats of an ACTFL/ILR-scale based test of Chinese speaking proficiency taken by 32 American students studying Chinese. The statistical analyses indicated that there was a consistent relationship between the ratings of the two test forms when there was only one rater; results with multiple raters were more problematic. However, the candidates overwhelmingly self-reported a preference for the live format (89%), describing the semi-direct version as more difficult and 'unfair' (cf. Hill 1998 mentioned below).

In another empirical study of the live OPI and the SOPI format, Clark and Hooshmand (1992) tested Arabic and Russian learners at the Defense Language Institute in both a face-to-face interview and one conducted via teleconferencing. Quantitative and questionnaire results suggested that the live format can be simulated in a teleconference and is acceptable to examinees as a substitute if necessary.

Stansfield and Kenyon's (1992) study also lends support to the equivalence of the OPI and a SOPI version. Their analyses showed that both measures are equally reliable and valid as measures of the same construct: 'they may be viewed as parallel tests delivered in two different formats' (p. 359). However, the SOPI may allow for a more accurate assessment of strategic competence, while the OPI is clearly preferable for tapping face-to-face interaction. And, as is now known, and has been demonstrated empirically, the same score on an OPI can represent different performances, and different scores can represent similar performances, due to the fact that a live interlocutor is present in the face-to-face interview.

At least two studies have investigated rater behaviour on the OPI. An early study by Shohamy (1983) examined the stability of oral assessment across

4 oral examination formats which differed by interviewer, speech style, and topic. Eighty-five Hebrew as a foreign language students in the U.S. were rated on these 4 methods by 2 independent raters; her analyses detected the main difference to be in the fourth test, where candidates reported information instead of being interviewed; she concludes that 'speech style and topic are significant factors influencing students' scores on oral proficiency' (p. 537). She suggests (somewhat contrary to her later opinion (Shohamy 1988)) that the OPI is well suited to testing other sorts of communicative behaviour.

Thompson (1995) also investigated interrater reliability on the OPI given to 795 candidates in 5 languages: English, French, German, Russian, and Spanish. A total of 175 raters assessed the interviews. Her results showed 'significant' overall interrater reliability with some variation due to proficiency level and language tested. Furthermore, she found that second ratings, done after the fact from audiotapes, were likely to be lower than original ratings.

Finally, Barnwell (1989) analyzed 4 OPIs in Spanish taken by American students and evaluated by 14 'naive' raters, all native speakers of Spanish, who were given OPI rating scales translated into Spanish. Barnwell found first, that the naive raters ranked the subjects in the same order, but the actual ratings for each of the 4 individual candidates varied, and second, that the naive raters were generally harsher than ACTFL trained raters.

Research on other oral examinations

There have also been many studies that have delved into these issues – validity, reliability, test method comparisons, and rating scale construction – with other oral examinations. Two early construct validation studies on the FSI (The Foreign Service Interview, the precursor to the ACTFL OPI) were conducted by Bachman and Palmer (1981, 1982). The first study (1981) examined the performance of 75 ESL students at an American university on 6 measures, comprised of 2 traits (speaking and reading) and 3 methods (interview, translation, and self-rating). Their results, based on correlations and factor analysis, showed respectable convergent and divergent validity for the FSI. In the second study, Bachman and Palmer (1982) used an adapted FSI oral interview procedure as one measure of communicative competence to assess the language ability of 116 ESL students at an American university. Factor analysis was employed to test three proposed traits (grammatical, pragmatic, and sociolinguistic competence) using the interview, a self-rating, a writing sample, and a multiple choice test. Their results suggested the existence of a general factor and two specific traits, grammatical/pragmatic competence and sociolinguistic competence.

The issue of rater reliability on other oral exams has been fruitfully explored as well. Several studies have explored the role of raters in the IELTS

Speaking Test (International English Language Testing System; UCLES 1999a). For example, Wylie (1993) probed the ability of raters to provide the same ratings, on two different occasions, of a single candidate performance on IELTS. Her examination of 18 Australian interviews showed high overall correlations (.906) for the ratings of all candidates. Styles (1993) also looked at rater behaviour on IELTS, specifically the reliability of ratings done in live assessments, from audiotapes, and from videotapes. He considered the assessments of 30 European candidates and concluded that the reliability of audiotaped assessments is as good as or better than videotaped assessments, both between and within raters, although the quality of the videorecordings was criticized by the raters and might have led to lower estimates of reliability. A somewhat contrary result was found by Conlan, Bardsley, and Martinson (1994), who compared live and audiotaped interviews of 27 IELTS candidates rated by 3 examiners. In 10 out of 27 cases, the audio recording was scored a full band lower than the live interview; they conclude that some examiners are more sensitive to extralinguistic, paralinguistic, and nonlinguistic information than others.

Bachman, Lynch, and Mason (1995) investigated the performance of 218 American Education Abroad students on a tape-mediated Spanish speaking test involving a summary of a lecture and an extended response. Both G-theory and FACETS were used to estimate rater reliability; they conclude that these two measurement models provide useful, complementary information: relative effects of facets are identified by G-theory while Rasch measurement allows the researcher to determine rater or task specific effects.

Brown (1995) examined a face-to-face oral test for Japanese tour guides for possible rater bias. Fifty-one subjects were assessed by 33 raters, including native and near-native speakers of Japanese who were either teachers of Japanese as a Foreign Language or actual tour guides. Her multifaceted Rasch results found no significant rating bias for either linguistic skill or task fulfillment, but the application of and perceptions about the specific rating criteria did differ among rater groups.

An interesting study of rater perceptions is McNamara and Lumley (1997). Using Rasch analysis to analyze the questionnaire responses from 7 Occupational English Test raters assessing the audiotapes of 70 candidates, they concluded that perceptions of poor audiotape quality led to harsher candidate ratings. Additionally, three salient factors emerged with respect to perceived competence of the interlocutor. First, there was a significant and consistent effect for candidates who were paired with less competent raters (they were rated higher, and thus compensated for poor interlocutor performance). And, a similar but stronger effect was detected for candidates who were paired with interlocutors who failed to achieve good rapport (again, they received higher ratings). They propose that rater perceptions of tape